in conjunction with the ACM International Symposium on High Performance Distributed Computing (HPDC)
June 21-25, 2010, Chicago, Illinois, USA

Emerging Computational Methods for the Life Sciences Workshop
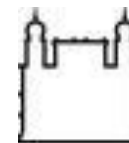
Data Parallelism in Bioinformatics Workflows Using Hydra

**Fábio Coutinho[1], Eduardo Ogasawara[1], Daniel de Oliveira[1] ,
Vanessa Braganholo[1] , Alexandre A. B. Lima[1] ,
Alberto M. R. Dávila[2] and Marta Mattoso[1]**

**[1]Federal University of Rio de Janeiro, Brazil
[2]FIOCRUZ, Brazil**

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

UFRJ
COPPE

Agenda

- Scientific experiments
- Blast workflow
- Data parallelization
- Hydra middleware
- Case study
- Measurements
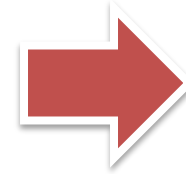- Provenance support
- Conclusion

Scientific Experiment Scenario

This scenario demands the execution of many programs as a chain of activities that may be represented as scientific workflows
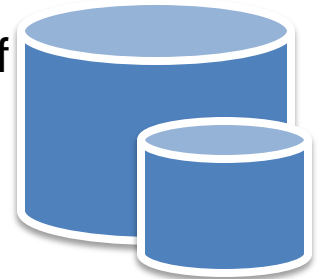
2. Data is first analyzed by some programs

1. Data is generated and collected

3. Large Volume of Data Produced ...

5. Final Results are analyzed

Program Y may demand HPC

4. ...which need to be processed by a cluster or a supercomputer…

Blast sequence analysis match

1. Genomics Data is collected.

Bioinformatics Tools

2. Fasta file is generated.

Fasta file

> ACBCD

Sequence Database

Ex. Genbank, DDBJ, EMBL, …

```
A C B C D B        A C - - B C D B
  | \ \ \          |       |   |
C A D B D        - C A D B - D -
```

3. Sequence alignment …

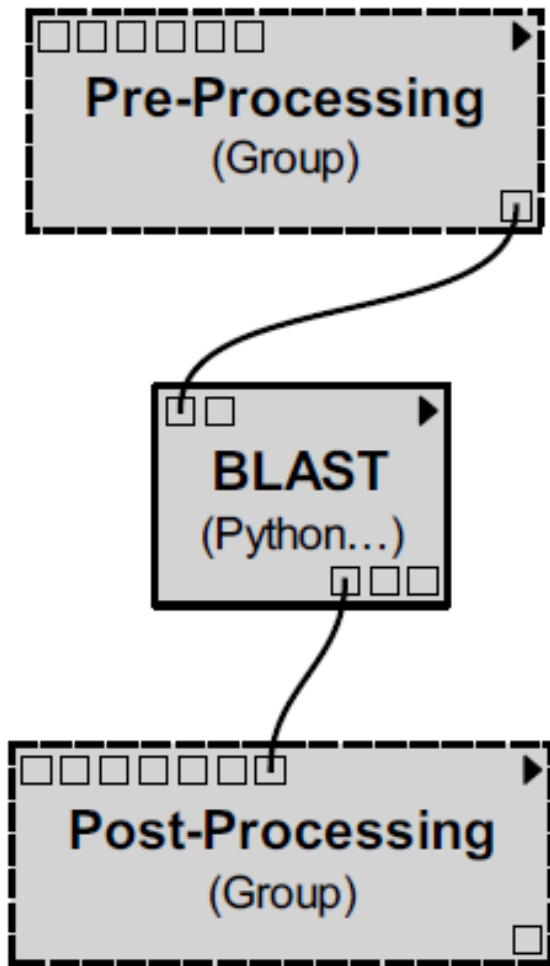4. ...which need to be processed In a HPC environment…

5. Final Results are analyzed

Provenance is a key to support the reproducibility of the experiment

Current Solutions

- mpiBlast, G-Blast and CloudBlast represent solutions for executing Blast in parallel on *cluster*, *grid* and *cloud* environments, respectively.
- BlastReduce: Using map-reduce approach to obtain data parallelization
  - These solutions are disconnected from the concept of scientific experiment and they are not concerned about capturing and analyzing provenance.
- Scientific Workflow Management Systems (SWfMS) is an alternative to represent a scientific experiment (Taverna Workflows, for example)
  - This solution brings some difficulties in obtaining parallelization

Experiments modeled as scientific workflows

Pre-Processing
(Group)

BLAST
(Python…)

Post-Processing
(Group)

Sequential Blast Workflow in VisTrails

- SWfMS is an alternative to represent a scientific experiment

- Obtaining parallelization and provenance gathering in distributed environments brings some challenges

Data Parallelization Difficulties

- Data fragmentation

  o Dependent on the data format

- Activity distribution in HPC from SWfMS

- Data aggregation
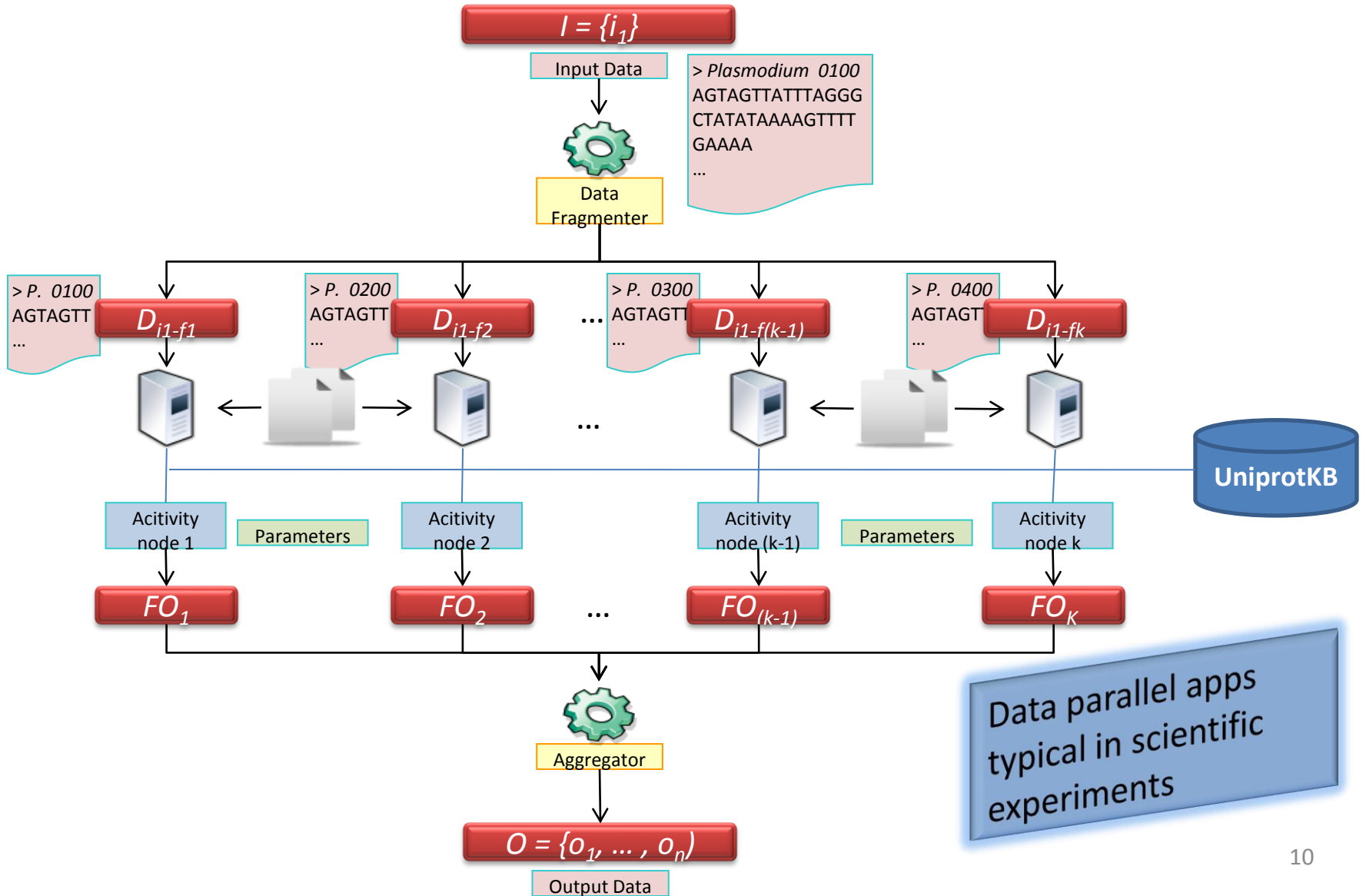
  o Dependent on the data format

FASTA Fragmentation

8
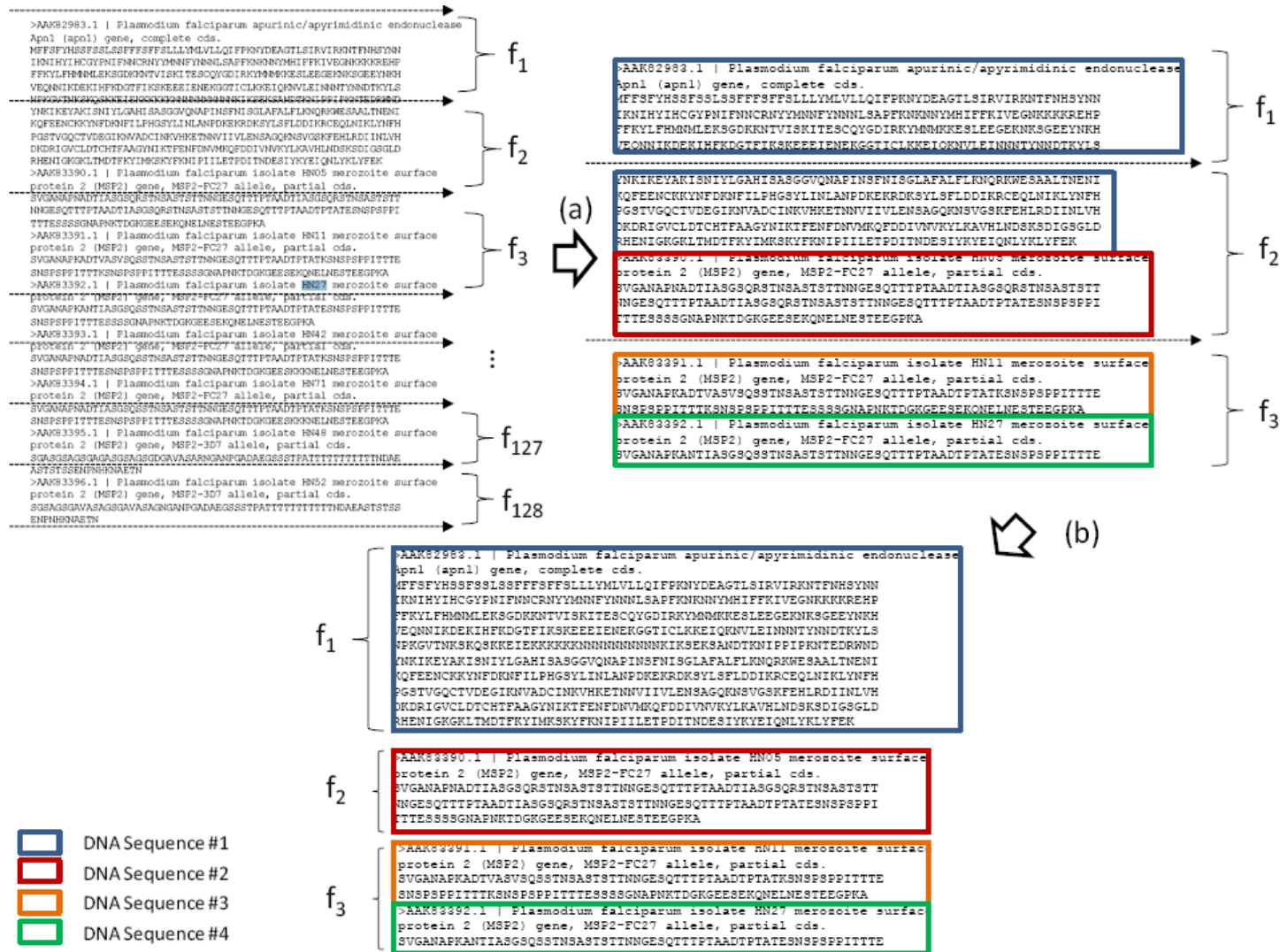
Types of FASTA data parallelization

	Sequence Fragmentation	Database Fragmentation
No data parallelization	No	No
Data parallelization	**Yes**	**No**
	No	Yes*
	Yes	Yes*

Data Parallelism

Hydra Middleware

- Middleware solution that bridges the SWfMS to HPC Environment supporting data parallelism

SWfMS

Hydra Middleware

MTC Environment

A set of components to be included on the workflow specification to control parallelism of activities as MTC

A set of components installed on the HPC environment that can register parallelism strategies to be reused with provenance gathering
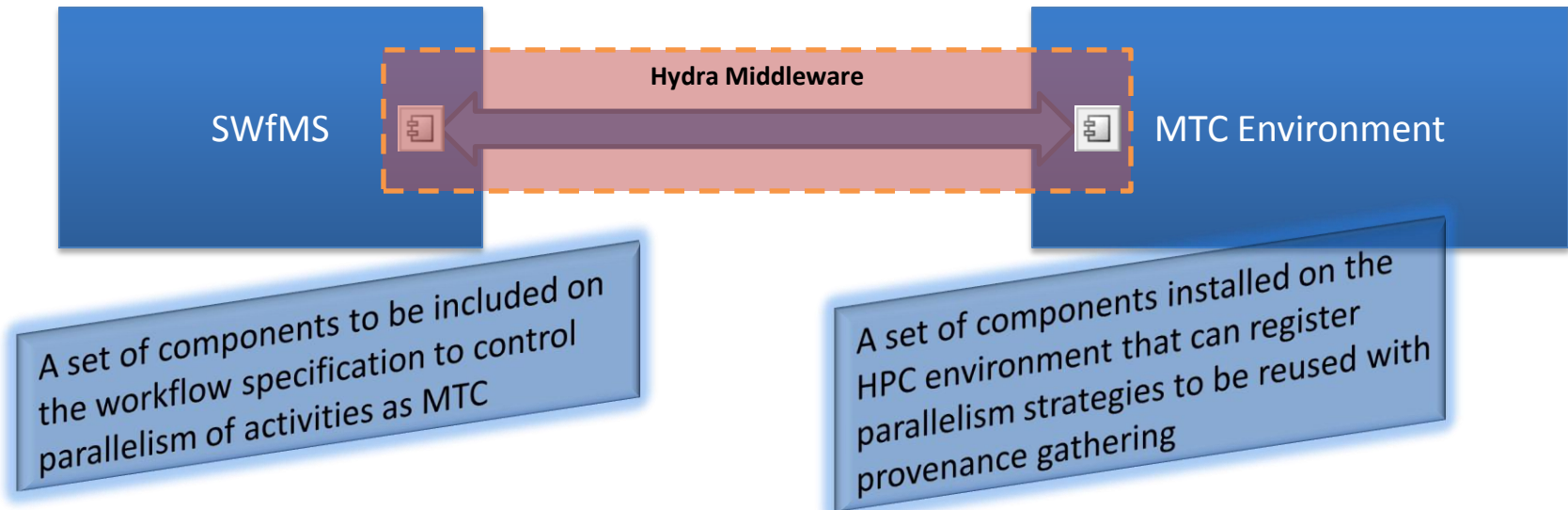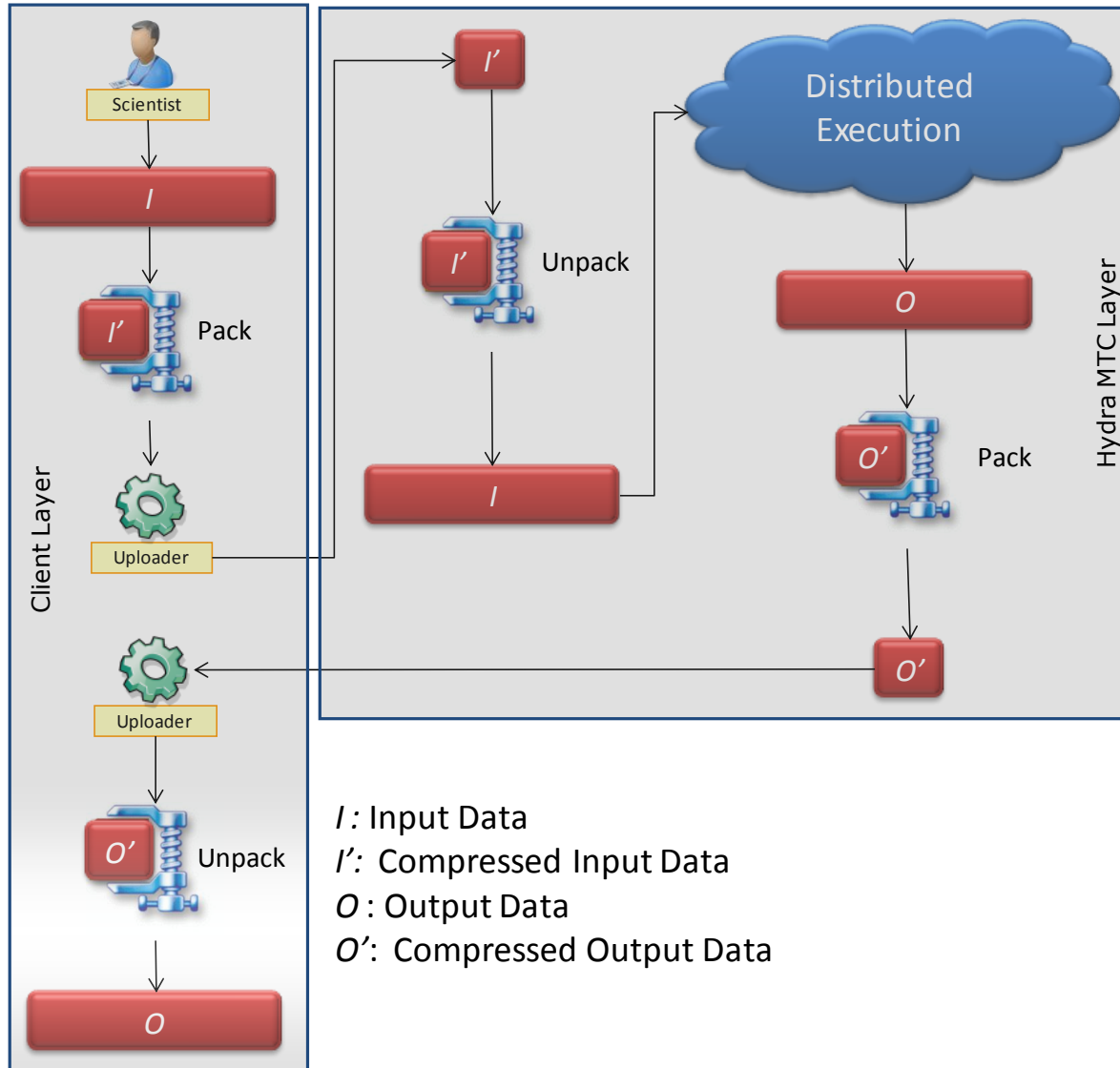
- Goal: reduce the complexity involved in designing and managing bioinformatics programs while collecting provenance data

Hydra Architecture
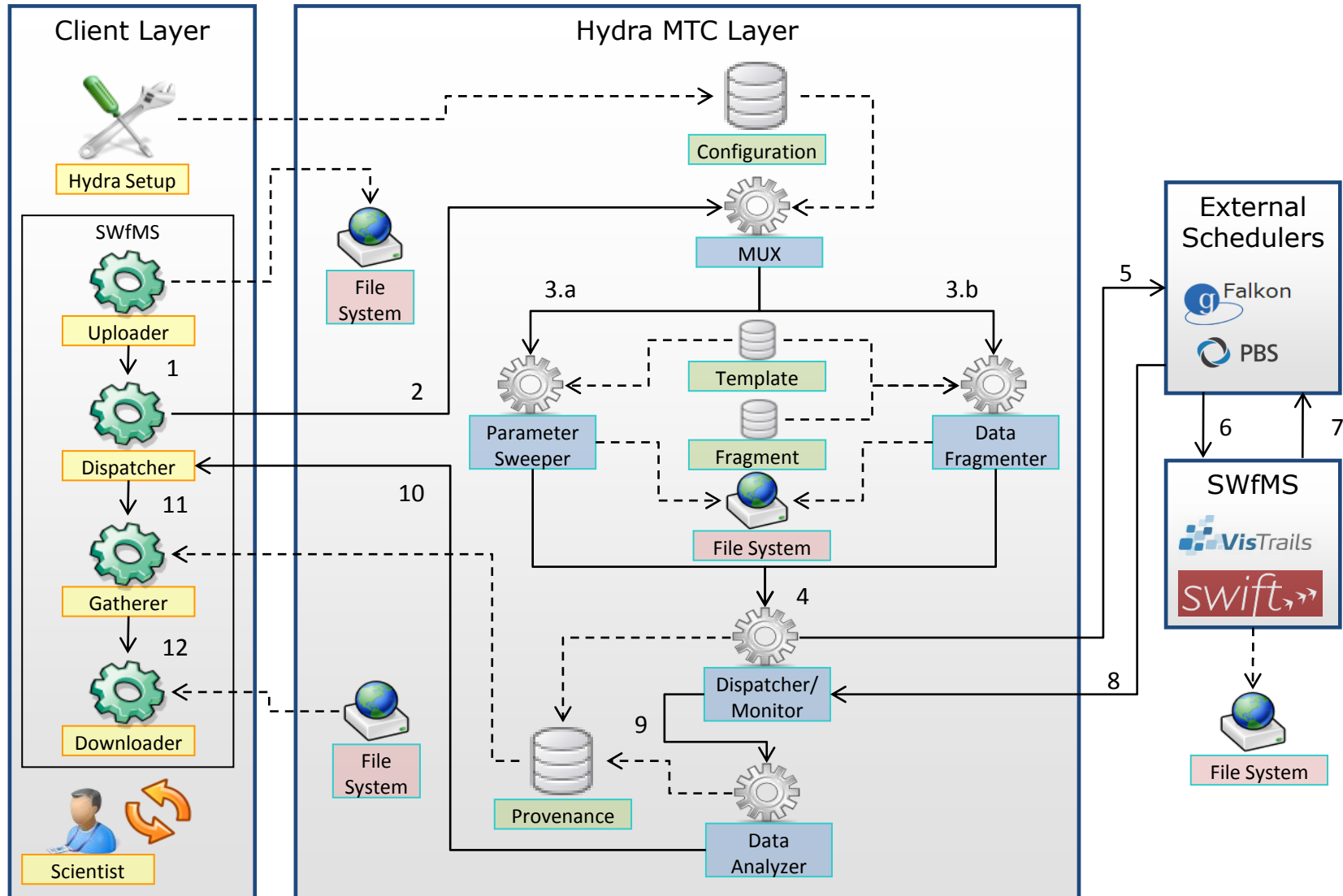
Compression Techniques in Hydra

I : Input Data
I' : Compressed Input Data
O : Output Data
O' : Compressed Output Data

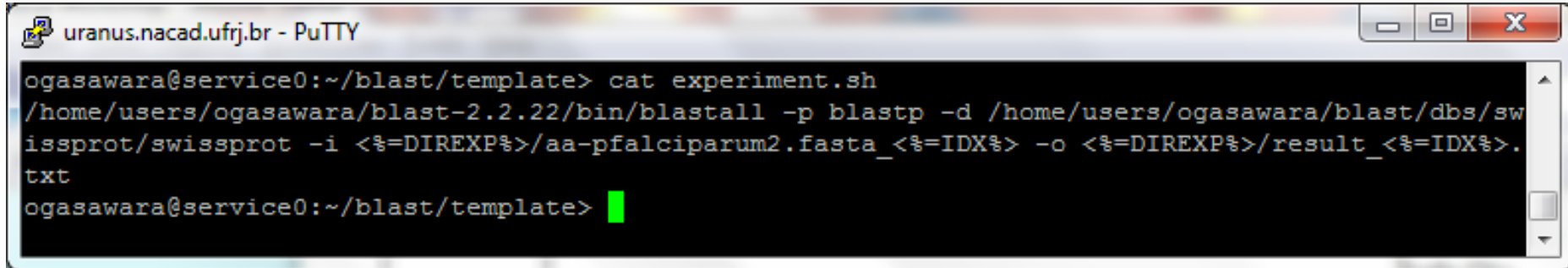Adding compression techniques to improve data transfer

13

Steps to use Hydra

- Adapt the workflow for distribution using Hydra
- Setup the data parallelization
  - Configure the workspace template and invoked program
- Setup a data fragmentation cartridge in Hydra
  - Develop a program that fragments a bioinformatics data file
- Setup a data aggregation cartridge in Hydra
  - A program that aggregate or merge data produced by individual workspaces

Blast workflow in VisTrails using Hydra

Hydra client components in a VT workflow
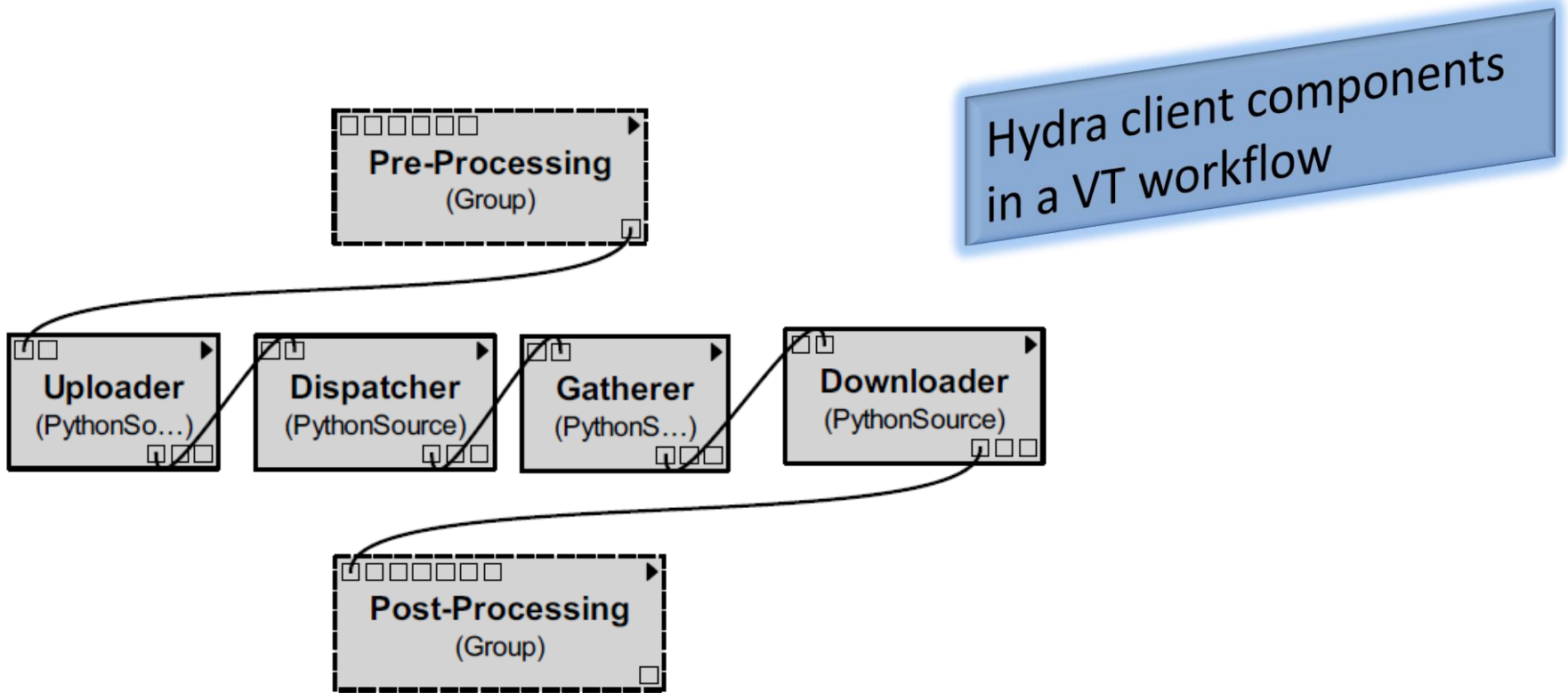
Setup the data parallelization

```
uranus.nacad.ufrj.br - PuTTY

ogasawara@service0:~/blast/template> cat experiment.sh
/home/users/ogasawara/blast-2.2.22/bin/blastall -p blastp -d /home/users/ogasawara/blast/dbs/sw
issprot/swissprot -i <%=DIREXP%>/aa-pfalciparum2.fasta_<%=IDX%> -o <%=DIREXP%>/result_<%=IDX%>.
txt
ogasawara@service0:~/blast/template>
```

Instrumenting the invocation of Blast

Case study

- BLAST tool
  - Identification of similar sequences between *Plasmodium falciparum* e o UniProtKB/SWISS-Prot database (june/2008)
 - Data fragmentation of input query

Experiment

- Workflow developed in VisTrails

- Hydra was setup with:
    - o Torque cartridge for job submission
    - o Data fragmentation cartridge using FASTASplitter
    - o Workspace configuration for blast invocation
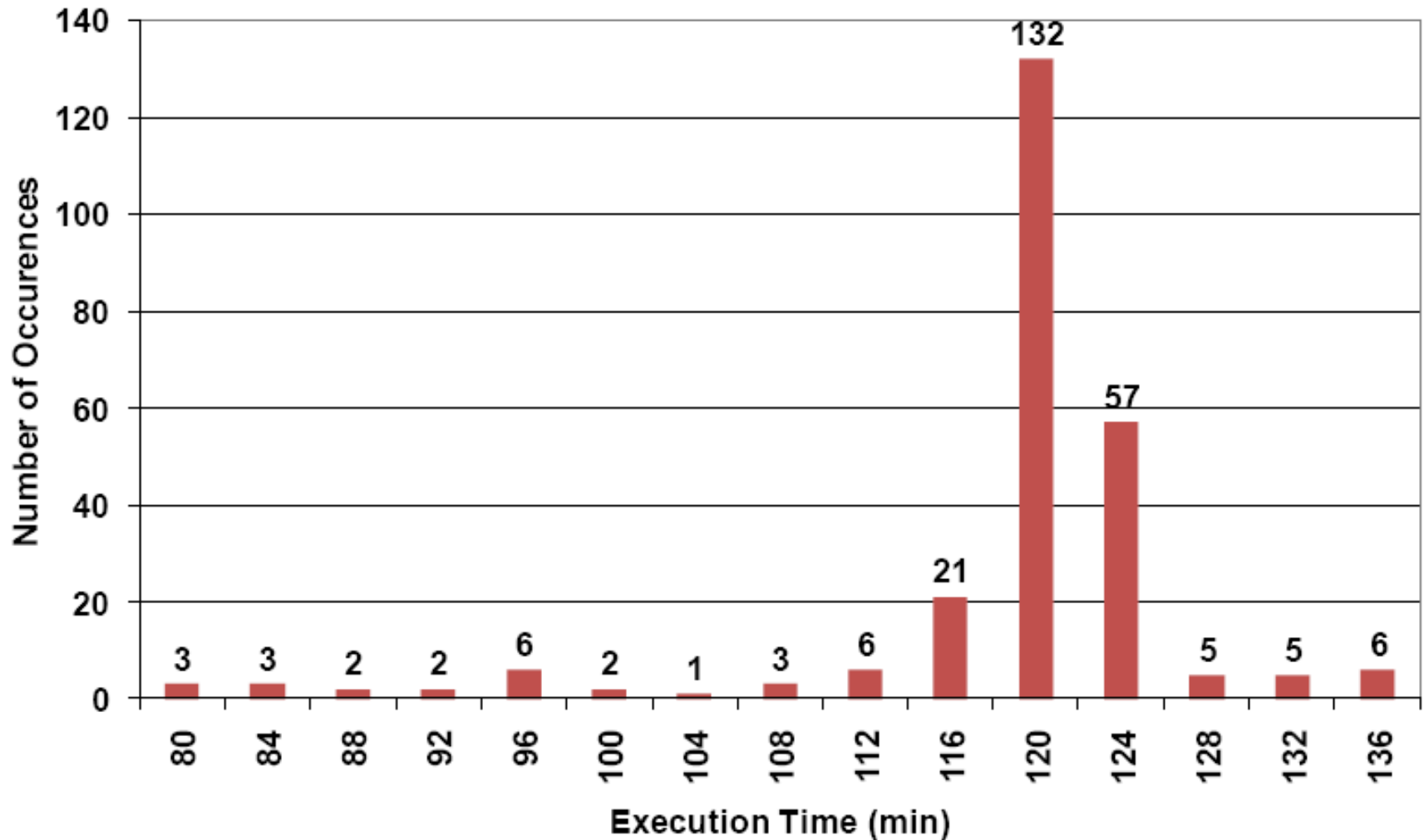   - o Data aggregation cartridge using BlastMerger

Experiment results

- Execution mode in 64 node Altix ICE 8200 with Xeon 8 cores per node installed at Nacad High Performance Computing Center at Federal University of Rio de Janeiro

- Speed-up measured

Workflow execution time
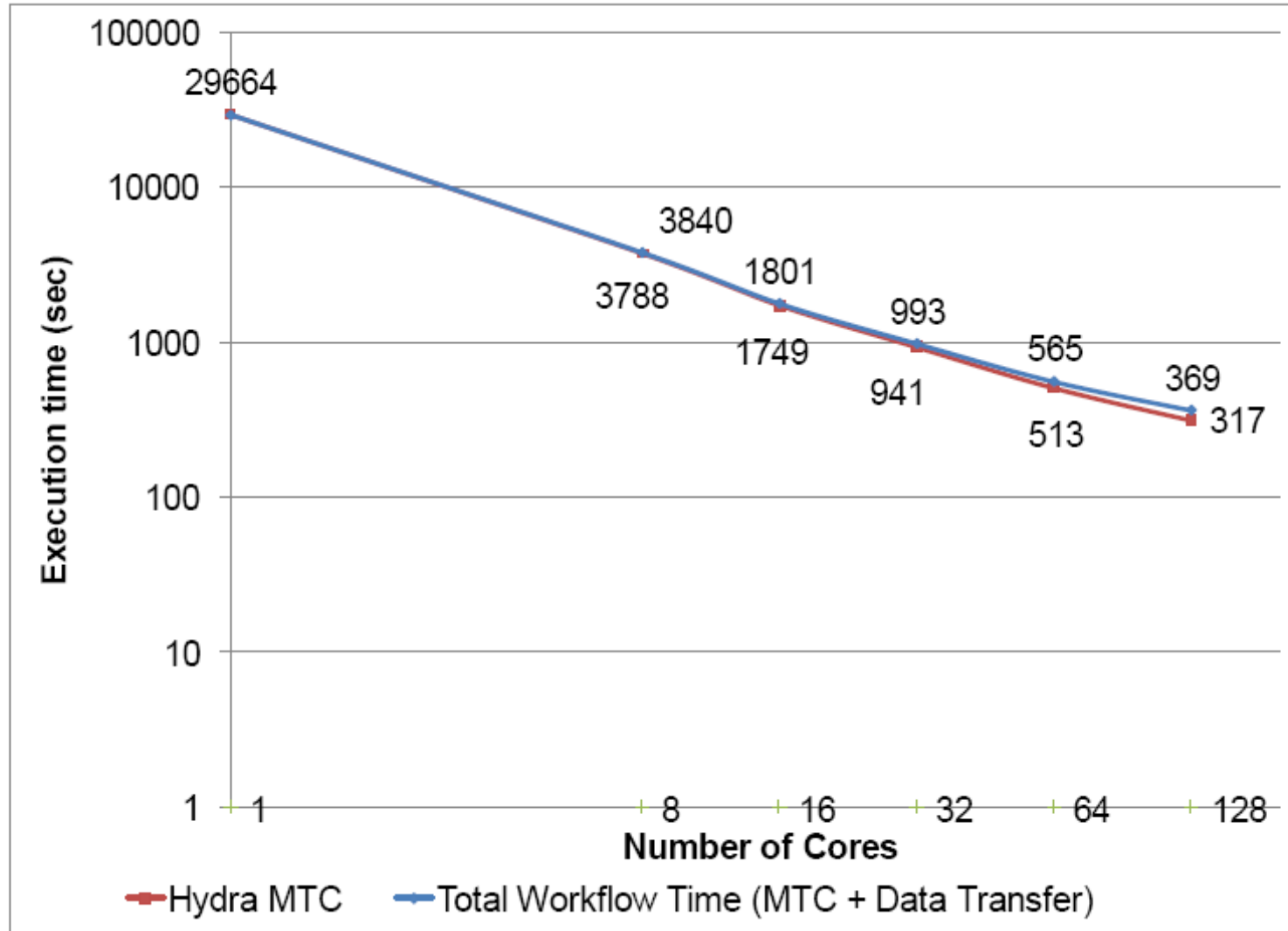
Activity execution time distribution

Provenance Support

o Where are the merged results for some particular experiment?

o What was the alignment method that lead to the best result?

o What were the parameters that lead the best result?

Provenance summary

Summary of a Hydra MTC execution using 4 nodes with 8 cores obtained from the provenance repository

Number of activities (fragments explored)	256
Start date/time	2010-03-09 12:40
Uploader time (compression + data transfer + decompression)	0.8 min
HPC Execution Time (Dispatcher + Gatherer)	941 min
Downloader time	0.1 min
Total Execution Time	**941.9 min**
Speedup	31.5

Conclusions

- Hydra can be a bridge the SWfMS and the HPC environment
  - Data parallelism using workflows
  - Evaluated in a real case bioinformatics experiment using BLAST with little overhead
  - Supports distributed provenance gathering
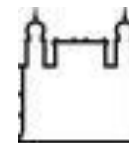 - Good speed-up and execution time

Future Work

- Balance workload between nodes
- New cartridges for data fragmentation and aggregation
- Automatic setup for fragment size and used nodes

in conjunction with the ACM International Symposium on High Performance Distributed Computing (HPDC)
June 21-25, 2010, Chicago, Illinois, USA

Emerging Computational Methods for the Life Sciences Workshop

Data Parallelism in Bioinformatics Workflows Using Hydra

Thank you!

**Fábio Coutinho, <u>Eduardo Ogasawara</u>, Daniel de Oliveira,
Vanessa Braganholo, Alexandre A. B. Lima,
Alberto M. R. Dávila and Marta Mattoso**

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

UFRJ COPPE